

Data-driven clustering of smart farming to optimize agricultural practices through machine learning

Pattharaporn Thongnim¹, Phaithoon Srinil², Thanaphon Phukseng²

¹Department of Mathematics, Faculty of Science, Burapha University, Chonburi, Thailand

²Information Technology and Data Science, Faculty of Science and Arts, Burapha University, Chanthaburi, Thailand

Article Info

Article history:

Received Sep 22, 2024

Revised Oct 22, 2024

Accepted Nov 19, 2024

Keywords:

Clustering

Data transformation

Log-transform

Machine learning

Smart farming

ABSTRACT

This study investigates the optimization of durian farming practices in Eastern Thailand using data-driven clustering techniques. The research aims to identify distinct agricultural patterns and improve resource allocation in durian production. K-means clustering is applied to durian production area and yield data from 2012 to 2023. Cluster quality is assessed using the Davies-Bouldin index (DBI), Dunn index, and Silhouette score. The methodology included comparing clustering results before and after log transformation of the data. Three main clusters are identified which are large-scale high-yield producers, small-scale lower-yield areas, and medium-scale producers with moderate yields. Notably, log transformation did not consistently improve clustering performance with original data often producing better-defined clusters. This finding highlights the importance of carefully considering data pre processing methods. Furthermore, the data-driven clustering offers valuable insights for precision agriculture by identifying regions with higher productivity allowing for targeted interventions and better resource allocation. The results can guide farmers in optimizing durian cultivation strategies, potentially leading to increased yields and more sustainable farming practices in Eastern Thailand's durian industry.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Thanaphon Phukseng

Information Technology and Data Science, Faculty of Science and Arts, Burapha University

Chanthaburi, Thailand

Email: thanaph@buu.ac.th

1. INTRODUCTION

Southeast Asia has a great variety of fruits such as mangoes, lychees, rambutans, and longans which grow well in the tropical climate [1], [2]. These fruits are important for local diets and also for their cultural value [3]. Growing and selling these fruits are key to the food security and economic stability of many Southeast Asian countries. For example, Thailand, Vietnam, Indonesia, Cambodia, and the Philippines are some of the top exporters of tropical fruits helping to feed people both locally and around the world [4], [5]. However, farmers in this region face big challenges. They need to produce more fruits to meet the increasing demand from both local and international markets but they also have to do this in a way that does not harm the environment. Climate change with more extreme weather, changes in rainfall and higher temperatures is already making it harder to grow fruits [6], [7]. To deal with these problems, farmers are starting to use more sustainable farming methods such as organic farming and growing crops that can handle the changing climate [8]. Moreover, fruits are not merely agricultural products but are deeply woven into the social and cultural fabric of the region [9].

Festivals celebrating the harvest of these fruits are common and they often play a symbolic role in religious and cultural rituals [10]. For example, bananas and oranges are frequently given as offerings in temples and during significant ceremonies showing their value beyond just food.

The growing global demand for tropical fruits has encouraged these countries to invest in better agricultural practices, infrastructure and technologies to maintain and enhance their competitive edge in the market. Governments and private sectors are increasingly focusing on improving supply chain efficiencies reducing post harvest losses and ensuring that their fruits meet international quality and safety standards. Therefore, the integration of technology especially artificial intelligence (AI), machine learning, and data science has played a significant role in transforming the agricultural sector [11]-[13]. These advanced technologies are being used to analyze vast amounts of data collected from farms such as soil quality, weather conditions, and crop health to make more informed decisions.

For example, AI driven predictive models can forecast crop yields and identify the best times for planting and harvesting helping farmers optimize their resources and increase productivity [14]. Machine learning algorithms are also being used to improve the efficiency of the supply chain by predicting demand patterns, optimizing transportation routes, and reducing waste [15]. By analyzing historical data and real time information, these technologies can help reduce post harvest losses and ensure that fruits reach markets at their peak quality [16]. Additionally, data science techniques are enabling better tracking and monitoring of fruits throughout the supply chain ensuring that they meet international quality and safety standards [17]. Therefore, this is global trade boosts the economies of these countries and promotes cultural exchange as people around the world become more familiar with and develop a taste for Southeast Asian fruits. The use of AI, machine learning, and data science in agriculture is helping these countries stay competitive in the global market while also contributing to the sustainability and efficiency of their farming practices. As a result, these technologies are supporting economic growth and helping to preserve the cultural and agricultural heritage of the region.

One significant issue is the digital divide where many farmers lack access to the necessary technology, internet connectivity and technical knowledge to fully benefit from these innovations [18]. The high cost of implementing advanced technology can also be prohibitive for smart farming limiting their ability to compete with larger agricultural enterprises [19]. Without proper support and infrastructure, the potential benefits of AI, machine learning, and data driven farming could be unevenly distributed, exacerbating existing inequalities in the agricultural sector. Additionally, integrating these technologies into traditional farming practices requires careful consideration of local cultural and environmental contexts to ensure that they are both effective and sustainable.

Therefore, this research focuses on analyzing durian farms in Eastern Thailand employing advanced machine learning and data science methods to uncover patterns and groupings among different provinces based on the area and production of these farms. Specifically, the study utilizes K-means clustering, a widely used machine learning algorithm to partition the data into clusters that share similar characteristics. The Elbow and Silhouette methods are applied to determine the optimal number of clusters ensuring that the analysis provides meaningful and well defined groupings. Additionally, the study employs performance metrics such as the Davies-Bouldin index (DBI) and Dunn index to evaluate the quality of the clusters formed. By integrating these techniques, the research aims to provide actionable insights that can optimize farming practices, enhance regional productivity, and support sustainable agricultural development in Eastern Thailand.

2. METHOD

2.1. Data collection

The data for this study is sourced from the Chanthaburi Data Center in Eastern Thailand covering durian farming activities from 2012 to 2023. The dataset includes key variables with the province and district which identify the specific geographic locations of the durian farms, the area of production in hectares which indicates the total land used for cultivation and durian yield in kilograms per hectare representing the productivity of the farms. This dataset offers a detailed overview of durian farming across various provinces allowing for an in depth examination of regional differences in farming practices and productivity.

2.2. Data cleaning and data transformation

Data cleaning is an important step in preparing a dataset for analysis, especially when dealing with missing values and outliers. To handle missing values, start by identifying any gaps in the data. Depending on how much data is missing and the importance of the missing values, replace them using the mean, median

and mode of the relevant column which helps keep the dataset accurate without losing too much information. However, if a lot of data is missing it could lead to errors, it might be better to remove the affected rows and columns altogether.

Additionally, identifying and addressing outliers is crucial as outlier data can significantly skew results. However, in this research, outliers were not removed from the dataset. This decision was made to study the model selection patterns under diverse data conditions and because the dataset is not particularly large. Retaining all data points, including outliers, allows for a more comprehensive view of the data landscape. Ensuring the dataset is clean, consistent, and complete will improve the accuracy and reliability of the subsequent analysis.

Feature scaling in the data transformation process is a crucial step, particularly when using clustering algorithms like K-means, which are sensitive to the scale of data. In this research, log transformation was applied as the primary method of data scaling. This approach, along with other normalization and standardization techniques, ensures that all variables contribute equally to the analysis, preventing any single feature from disproportionately influencing the clustering results. For instance, in the context of clustering durian farms, features such as the area of production was log-transformed. This scaling method helps to compress the range of large values and expand the range of small values addressing skewness in the data distribution. It ensures that differences in the units and ranges of these features do not skew the clustering process leading to more accurate and meaningful groupings of the data. The use of log transformation is particularly effective for variables with a wide range and those that follow a multiplicative rather than additive pattern.

2.3. Clustering analysis

The core of this research involves the application of clustering algorithms to group provinces based on the characteristics of their durian farms. The following steps outline the clustering process.

2.3.1. Model clustering

The K-means algorithm was chosen due to its effectiveness in partitioning data into distinct clusters based on similarity. The algorithm works by minimizing the within cluster sum of squares and iteratively adjusting cluster centroids to improve clustering quality [20], [21]:

$$WCSS = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2,$$

where $WCSS$ is the objective function to be minimized, k is the number of clusters, n is the number of observations, $x_i^{(j)}$ is the i -th observation belonging to the j -th cluster and c_j is the centroid of the j -th cluster. The algorithm proceeds by alternating between two steps:

– Assignment step:

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \quad \forall j, 1 \leq j \leq k\},$$

– Update step:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j,$$

where $S_i^{(t)}$ is the set of points assigned to cluster i at iteration t and $m_i^{(t)}$ is the mean of points in $S_i^{(t)}$. The optimal number of clusters (k) was determined using the Elbow method [22], [23]. The Elbow method involves plotting $WCSS$ against the number of clusters (k) and looking for the Elbow point. This point can be mathematically defined as the point of maximum curvature on the $WCSS$ curve.

$$\frac{d^2 WCSS}{dk^2}.$$

The k value at which this second derivative is maximized can be considered the optimal number of clusters. Alternatively, the percentage of variance explained can be used:

$$\text{Variance explained} = 1 - \frac{WCSS_k}{WCSS_1},$$

where $WCSS_k$ is the within-cluster sum of squares for k clusters, and $WCSS_1$ is the $WCSS$ when $k=1$.

2.3.2. Clustering validation

To validate the quality of the clusters obtained in this study, three well-known performance metrics are employed such as the DBI [24], the Dunn index [25], and the Silhouette score [26]. Each of these metrics provides a unique perspective on the separation and compactness of the clusters which are essential characteristics of well formed clusters.

- DBI: a lower DBI indicates better cluster separation and compactness. It was used to assess the average similarity ratio of each cluster with its most similar cluster.

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + S_j}{d_{ij}} \right),$$

where S_i is the average distance between each point in cluster i and the centroid of cluster i , d_{ij} is the distance between the centroids of clusters i and j and k is the total number of clusters.

- Dunn index: this index measures the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance. Higher Dunn index values indicate well-separated and compact clusters.

$$\text{Dunn index} = \frac{\min_{1 \leq i < j \leq k} d(C_i, C_j)}{\max_{1 \leq i \leq k} \delta(C_i)},$$

where $d(C_i, C_j)$ is the distance between clusters C_i and C_j , $\delta(C_i)$ is the maximum distance between points within cluster C_i and k is the number of clusters.

- Silhouette score: this metric measures how similar each data point is to its own cluster compared to other clusters. Higher Silhouette scores indicate well-defined clusters. The k -value that maximized the Silhouette score was considered optimal.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average distance between point i and other points in the same cluster and $b(i)$ is the minimum average distance from point i to points in the nearest different cluster.

2.4. Data visualization

To gain deeper insights into the clustering results, data visualization techniques are employed. Visualizing the data distribution is crucial to understanding the characteristics of the dataset before and after the clustering process. In this study, histograms and box plots [27] are used to display the distribution of key variables such as durian production area and yield across the different provinces. Histograms help identify patterns, trends, and outliers within the data while box plots provide a clear summary of the central tendency, variability and potential outliers for each variable. These visualizations allow for a comparison of the distribution of variables between different clusters and offer insights into how the data is spread across the provinces. By examining the variability and identifying outliers, it becomes easier to evaluate the effectiveness of the clustering model and ensure that the clusters represent meaningful groupings based on the underlying data.

2.5. Interpretation of results

The clustering results are analyzed to identify key patterns and insights.

- Cluster characteristics: each cluster was analyzed to determine the common characteristics shared by provinces within the same cluster, such as similar yields, farming practices, or environmental conditions.
- Actionable insights: based on the identified clusters, recommendations were made for optimizing agricultural practices in each cluster. This included suggestions for resource allocation, technological interventions, and sustainable farming practices tailored to the specific needs of each cluster.

3. RESULT AND DISCUSSION

Figure 1 illustrates the distribution of average durian production area and yield for each province across multiple years. The box plots highlight the central tendency and variability within each province with outliers indicated beyond the whiskers. The left panel represents the average area of production (in hectares) showing the range of land use across provinces while the right panel represents the average yield (in kilograms per hectare) illustrating the productivity differences among provinces. The box plots provide insights into regional variations helping to identify provinces with consistently high or low production and yield.

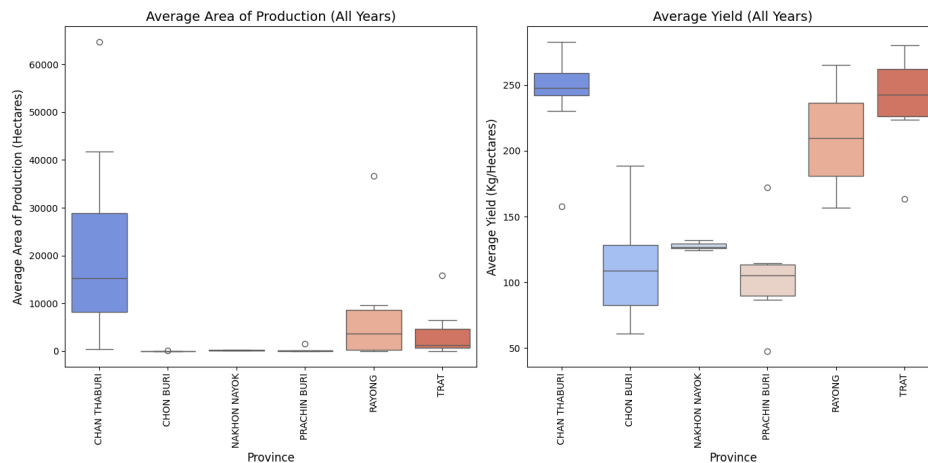


Figure 1. Box plot of average area of production and yield for durian farms across provinces (all years)

Figure 2 showcases a comparison between the original and log-transformed distribution of the durian production area. The original distribution, shown on the left, reveals significant skewness with a few provinces having much larger production areas compared to others. This skewness can create challenges in clustering analysis, as larger values can dominate and skew the results, potentially leading to less accurate clusters. The original data presents an uneven distribution of production areas, making it difficult for the clustering algorithm to group provinces meaningfully.

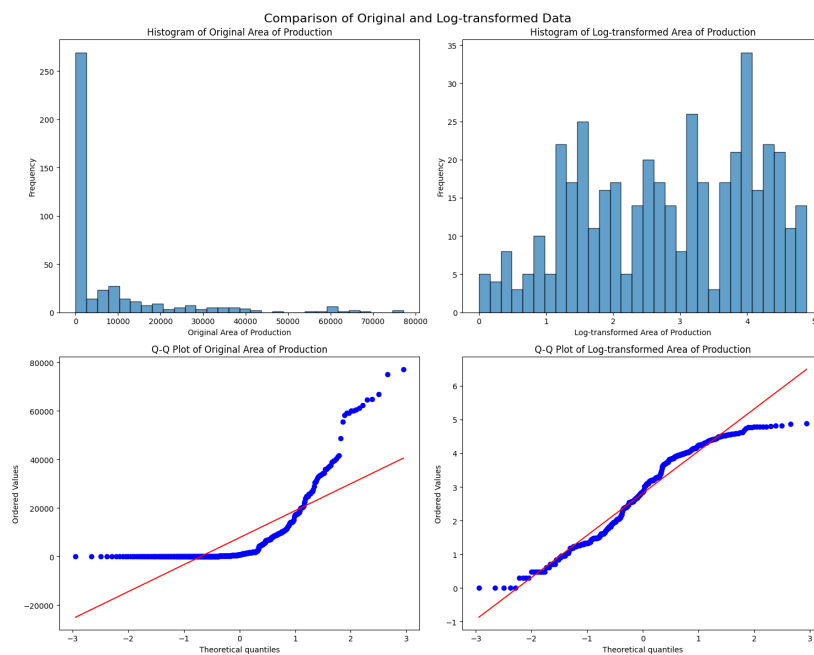


Figure 2. Comparison of distribution characteristics in original and log-transformed area of production

On the right, the log-transformed distribution addresses this issue by reducing the impact of extreme values. The log transformation compresses the range of large production areas and expands the range of smaller ones, resulting in a more balanced distribution. This normalization process helps improve the clustering performance by ensuring that no single province disproportionately influences the results. As a result, the log-transformed data allows for a more accurate representation of patterns and relationships among provinces in terms of durian production, enabling better-defined and more meaningful clusters.

Figure 3 illustrates the comparison between the original and log-transformed distribution of durian yield per hectare. The original distribution, shown on the left, demonstrates skewness, where a few provinces exhibit significantly higher yields compared to others. This uneven distribution suggests that some regions produce much more per hectare, while others lag behind, creating challenges for clustering algorithms to accurately group similar provinces. In this original form, the data is heavily influenced by outliers, which could result in clusters dominated by these extreme values, skewing the analysis and masking underlying patterns.

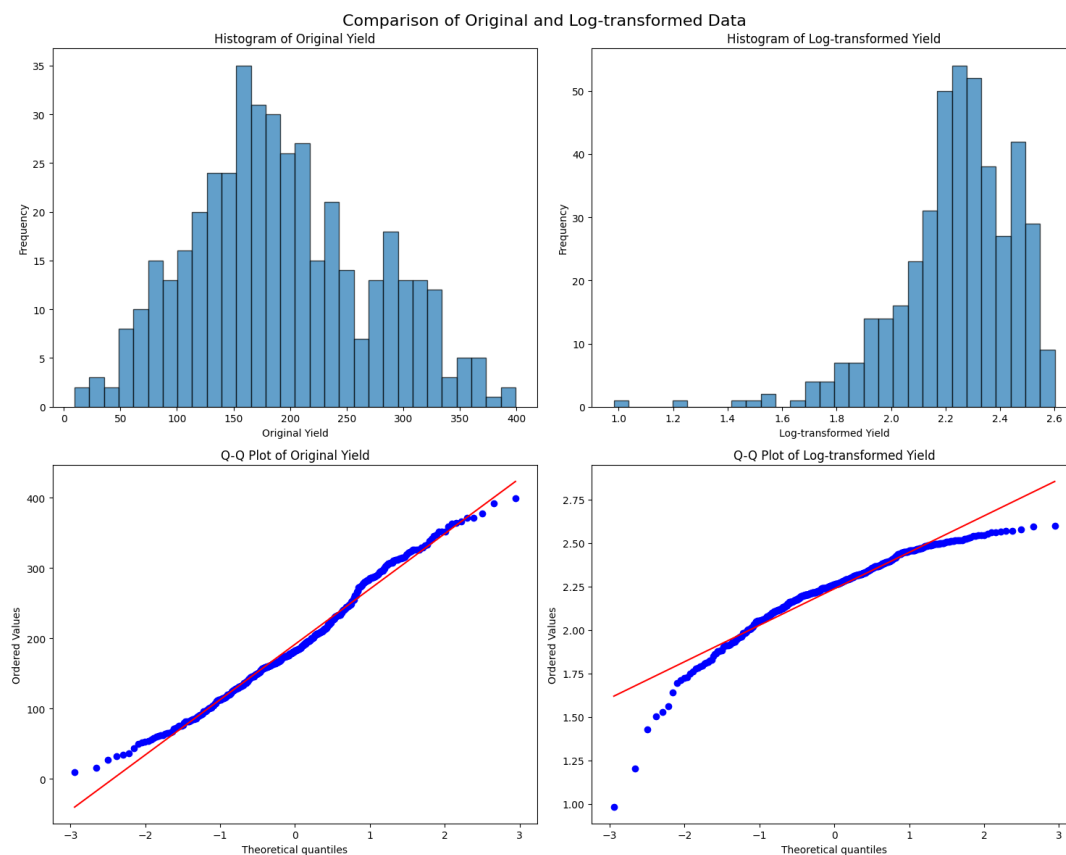


Figure 3. Comparison of distribution characteristics in original and log-transformed yield

On the right, the log-transformed distribution adjusts the skewness by compressing large yield values and expanding smaller ones. However, it introduces certain irregularities, such as a left-skewed tendency and bimodal patterns. Although the log transformation reduces the impact of extreme values, its effectiveness in aiding analysis and clustering depends on the dataset's structure. In this case, the transformation does not consistently enhance the accuracy of reflecting variations in durian yield across provinces.

Figure 4 illustrates the clustering of provinces from 2012 to 2023 based on durian production area and yield. Each province is assigned to one of three clusters, represented by different colors, indicating similarities in farming practices, land use, and productivity. The spatial distribution of clusters highlights regional differences in durian farming, with some provinces consistently grouped together over the years due to similar agricultural characteristics. This visualization provides a clear geographic perspective of the clustering results and allows for the identification of patterns and trends across the regions.

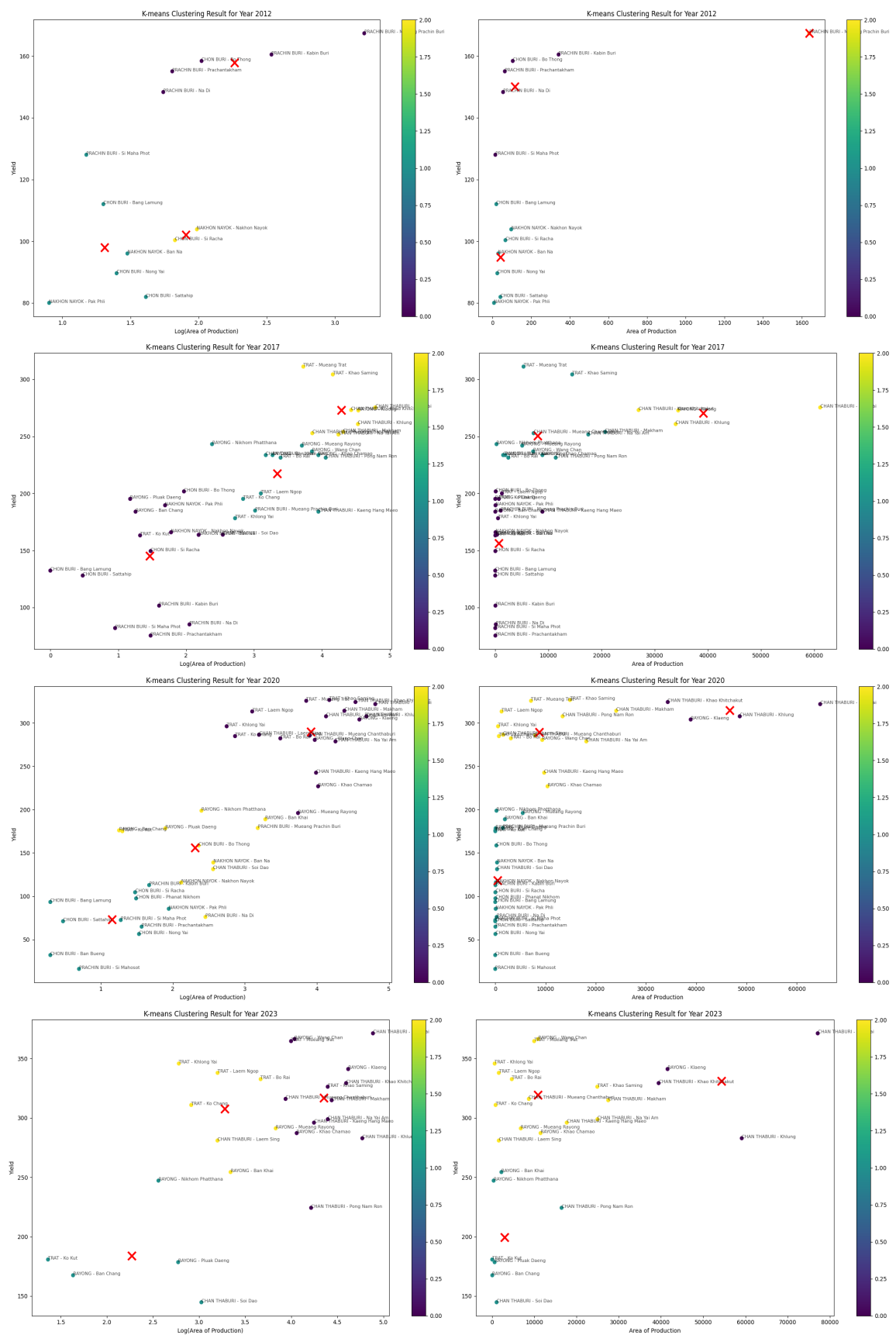


Figure 4. Cluster assignment of provinces (2012, 2017, 2020, and 2023)

The figure also highlights temporal changes in provincial clustering offering a deeper understanding of how the agricultural landscape may have shifted over time. This helps to identify provinces that have either increased and decreased in productivity and land use, thereby allowing researchers and policymakers to monitor the evolution of durian farming regions. The geographic distribution of the clusters can further inform strategic planning for resource allocation, technological interventions and the implementation of tailored farming practices to optimize yield and sustainability across Eastern Thailand.

The comparison of clustering metrics before and after log transformation reveals significant insights into the dataset's structure and the effectiveness of the transformation (Table 1). Overall, the log transformation generally did not improve clustering performance across the years studied. Davies-Bouldin Scores consistently increased after transformation, indicating potentially poorer cluster separation. Silhouette scores mostly decreased suggesting that the original data may have had better defined clusters. The Dunn index showed mixed results with some years improving and others deteriorating after transformation.

Table 1. Comparison of clustering metrics before and after log transform

Year	Davies-Bouldin Score		Dunn index		Silhouette score	
	Before	After	Before	After	Before	After
2012	0.302665	0.685463	0.396260	0.236223	0.615334	0.320194
2013	0.543252	0.653921	0.078082	0.070662	0.558279	0.489009
2014	0.593647	0.856772	0.036246	0.164710	0.520202	0.479697
2015	0.684264	0.904941	0.056896	0.162591	0.441512	0.446666
2016	0.586815	0.753699	0.098682	0.161728	0.521381	0.476560
2017	0.652970	0.742797	0.199246	0.101041	0.491824	0.407584
2018	0.634182	0.740250	0.070140	0.145778	0.479389	0.439386
2019	0.632832	0.716224	0.027092	0.124827	0.480100	0.486832
2020	0.507760	0.748040	0.220983	0.210851	0.597098	0.482537
2021	0.497913	0.741959	0.116295	0.122217	0.607284	0.475719
2022	0.528537	0.651915	0.067940	0.154245	0.625088	0.506990
2023	0.659915	0.888091	0.207901	0.105373	0.487491	0.354306

Among the years analyzed, 2022 stands out as having the most robust clustering results. It demonstrated the best Silhouette score both before (0.625088) and after (0.506990) transformation as well as the best Davies-Bouldin Score after transformation (0.651915). The year 2012 also showed notable performance particularly before transformation with the best Davies-Bouldin Score (0.302665) and Dunn index (0.396260). However, its performance declined more significantly after transformation compared to 2022, especially in terms of Silhouette score.

The varying impact of log transformation across different years suggests that the underlying data structure may be changing over time. This observation highlights the importance of carefully considering data transformation techniques in clustering analysis. While log transformation is often used to handle skewed data. In this case, it appears that the non transformed data yielded better-defined clusters in most instances. These findings emphasize the need for a thorough understanding of the dataset's characteristics and the potential effects of transformations when performing cluster analysis as the optimal approach may vary depending on the specific year and subset of data being analyzed.

Table 2 presents the summary statistics for the clusters generated based on the durian production area and yield in 2023. The table shows the mean, minimum and maximum values for both the area of production (in hectares) and the yield (in kilograms per hectare) for each of the three identified clusters. Cluster 0 which consists of provinces with the largest production areas has a mean production area of 54,317.25 hectares with yields averaging 331.12 kg/ha. This suggests that provinces in this cluster are the major producers of durian, both in terms of land area and productivity.

Table 2. Cluster statistics for year 2023

Cluster	Area of production			Yield		
	Mean	Min	Max	Mean	Min	Max
0	54,317.25	39,444	77,130	331.12	282.88	371.20
1	2,963.14	23	16,457	199.66	144.80	254.40
2	10,869.00	604	27,508	319.26	280.96	366.40

In contrast, Cluster 1 with a mean production area of only 2,963.14 hectares represents provinces

with much smaller durian farming areas. The average yield in this cluster is lower at 199.66 kg/ha indicating that these provinces have less land dedicated to durian farming and lower productivity per hectare. Cluster 2 with a mean production area of 10,869 hectares and an average yield of 319.26 kg/ha represents provinces with moderate levels of production and relatively high yields. These insights provide a clear segmentation of durian farming regions with Cluster 0 representing large scale with high yield producers while Clusters 1 and 2 represent smaller and moderate scale producers with varying levels of productivity.

Table 3 shows the cluster assignments for individual provinces and districts based on the durian production data from 2023. Each row represents a province and its respective district, categorized into one of the three clusters identified in the analysis. The clusters are differentiated by shared characteristics such as production area and yield with Cluster 0 typically representing provinces with large production areas and high yields while Clusters 1 and 2 represent provinces with smaller production areas and varying productivity levels.

Table 3. Cluster assignments for provinces and districts in 2023

Province	District	Cluster
Chanthaburi	Khao Khitchakut	0
Chanthaburi	Khlung	0
Chanthaburi	Tha Mai	0
Rayong	Klaeng	0
Chanthaburi	Pong Nam Ron	1
Chanthaburi	Soi Dao	1
Rayong	Ban Chang	1
Rayong	Ban Khai	1
Rayong	Nikhom Phatthana	1
Rayong	Pluak Daeng	1
Trat	Ko Kut	1
Chanthaburi	Kaeng Hang Maeo	2
Chanthaburi	Laem Sing	2
Chanthaburi	Makham	2
Chanthaburi	Mueang Chanthaburi	2
Chanthaburi	Na Yai Am	2
Rayong	Khao Chamao	2
Rayong	Mueang Rayong	2
Rayong	Wang Chan	2
Trat	Bo Rai	2
Trat	Khao Saming	2
Trat	Khlung Yai	2
Trat	Ko Chang	2
Trat	Laem Ngop	2
Trat	Mueang Trat	2

This table provides a detailed breakdown of the geographic distribution of the clusters within the provinces of Chanthaburi, Rayong, and Trat. For example, districts such as Khao Khitchakut and Khlung in Chanthaburi are grouped into Cluster 0 indicating they are major durian production areas. In contrast, districts such as Pong Nam Ron in Chanthaburi and Ban Chang in Rayong fall into Cluster 1 which corresponds to regions with smaller production areas and relatively lower yields. This level of detail facilitates a more granular understanding of how different regions contribute to overall durian production and how they vary in terms of agricultural efficiency and land use.

In this study, the clustering analysis provided valuable insights into the patterns and groupings of durian farms based on production area and yield across provinces in Eastern Thailand. The application of K-means clustering, validated by multiple performance metrics such as the DBI, Dunn index, and Silhouette score allowed for the identification of well defined clusters that reflect the diversity in agricultural practices and productivity levels among the regions. The analysis showed that log transformation while often useful for normalizing skewed data did not consistently improve clustering performance across all years with some years showing better defined clusters in the original dataset. The identification of distinct clusters offers practical implications for optimizing farming practices by highlighting key differences between regions with large-scale production and those with lower yields. These findings provide a data-driven foundation for enhancing durian cultivation practices through targeted resource allocation and the adoption of tailored smart farming techniques.

The clustering analysis of durian production in Eastern Thailand for 2023 reveals distinct patterns

across the three main provinces which is Chanthaburi, Rayong, and Trat. This analysis based on production area size and yield, categorizes districts into three clusters providing valuable insights into the region's agricultural landscape.

Chanthaburi exhibits the most diverse clustering with districts spread across all three clusters. The province's major durian producing areas including Khao Khitchakut, Khlung, and Tha Mai, fall into Cluster 0, characterized by large production areas (averaging 54,317.25 hectares) and high yields (averaging 331.12 kg/hectare). The majority of Chanthaburi's districts such as Kaeng Hang Maeo, Laem Sing, Makhm, Mueang Chanthaburi, and Na Yai Am, belong to Cluster 2, featuring medium sized production areas (averaging 10,869 hectares) and relatively high yields (averaging 319.26 kg/hectare). A few districts, including Pong Nam Ron and Soi Dao, fall into Cluster 1 with smaller production areas (averaging 2,963.14 hectares) and lower yields (averaging 199.66 kg/hectare). Rayong shows a different pattern with most districts in Clusters 1 and 2. Only Klaeng district is in Cluster 0, while Ban Chang, Ban Khai, Nikhom Phatthana, and Pluak Daeng are in Cluster 1. Khao Chamao, Mueang Rayong, and Wang Chan districts comprise Rayong's Cluster 2 representation. Trat presents a more homogeneous picture, with all its districts except one in Cluster 2. These include Bo Rai, Khao Saming, Khlong Yai, Ko Chang, Laem Ngop, and Mueang Trat. The exception is Ko Kut district, which falls into Cluster 1.

This clustering analysis provides crucial insights for precision agriculture planning and development in Eastern Thailand. It highlights areas of high productivity such as the Cluster 0 districts in Chanthaburi and Rayong which may serve as models for best practices. Conversely, Cluster 1 areas found across all three provinces may benefit from targeted interventions to improve yield and expand production. The prevalence of Cluster 2 districts particularly in Trat and parts of Chanthaburi suggests a stable mid-range production capacity with potential for optimization. By understanding these clusters, agricultural planners and policymakers can tailor their approaches to each area's specific characteristics, potentially leading to more efficient resource allocation and improved overall durian production in the region.

4. CONCLUSION

This research showed how data-driven clustering methods can be used to analyze and improve durian farming practices in Eastern Thailand. By using K-means clustering on data about durian production areas and yields, distinct groups of provinces with similar farming characteristics are identified. Measures like the DBI, Dunn index, and Silhouette score ensured that the clusters were clear and meaningful. Log transformation is applied to adjust the data, but the results varied each year, highlighting the importance of understanding the data before making changes.

The results of this study provide useful ideas for improving durian farming through precision agriculture. By identifying areas with higher productivity and making better use of resources, farmers can increase overall yield efficiency. The clusters found in this research can also guide specific actions such as the use of smart farming technologies designed to meet the needs of each region. These insights can help durian farming remain sustainable despite challenges such as climate change and shifting market demands.

In future work, additional variables could be incorporated to create more comprehensive models of durian farming efficiency. Exploring more advanced machine learning models beyond K-means, such as hierarchical clustering and deep learning techniques could further improve the accuracy of groupings. Additionally, refining the data preprocessing steps including more robust handling of outliers and testing different transformation techniques would enhance the overall quality of the clustering process.

ACKNOWLEDGMENTS




We would like to express our gratitude to the Regional Office of Agricultural Economics 6 for their invaluable support in providing comprehensive agricultural data for the Eastern region of Thailand. Their cooperation and willingness to share crucial information have significantly contributed to the depth and accuracy of this research. Additionally, we gratefully acknowledge the Data Center of Burapha University Chanthaburi Campus for providing a research grant and facilitating resources which have been instrumental in conducting this study.

REFERENCES




- [1] S. K. Mitra, "Overview of lychee production in the asia-pacific region," in *Lychee Production in the Asia Pacific Region. Food and Agricultural Organization of the United Nations, Bangkok, Thailand*, pp. 5–13, 2002.
- [2] S. Salakpetch, "An overview of tropical fruit production in Thailand," *Hawaii Tropical Fruit Growers*, vol. 45, no. 1, pp. 6–9, 2005.
- [3] J. M. L. Montesclaros and P. S. Teng, "Agriculture and food security in asia," *Climate Change, Disaster Risks, and Human Security: Asian Experience and Perspectives*, pp. 137–168, 2021, doi: 10.1007/978-981-15-8852-5_7.
- [4] Z. Mohamed, I. AbdLatif, and A. M. Abdullah, "E1-conomic importance of tropical and subtropical fruits," *Postharvest Biology and Technology of Tropical and Subtropical Fruits*, pp. 1–20, 2011, doi: 10.1533/9780857093622.1.
- [5] N. M. N. Rozana, C. Suntharalingam, and M. F. Othman, "Competitiveness of malaysia's fruits in the global market: Revealed comparative advantage analysis," *Malaysian Journal of Mathematical Sciences*, vol. 11, pp. 143–157, 2017.
- [6] P. Bhattacharjee, O. Warang, S. Das, and S. Das, "Impact of climate change on fruit crops-a review," *Current World Environment*, vol. 17, no. 2, p. 319, 2022, doi: 10.12944/CWE.17.2.4.
- [7] A. Shahzad *et al.*, "Nexus on climate change: Agriculture and possible solution to cope future climate change stresses," *Environmental Science and Pollution Research*, vol. 28, pp. 14211–14232, 2021, doi: 10.1007/s11356-021-12649-8.
- [8] R. Soni, R. Gupta, P. Agarwal, and R. Mishra, "Organic farming: A sustainable agricultural practice," *Vantage: Journal of Thematic Analysis*, vol. 3, no. 1, pp. 21–44, 2022, doi: 10.52253/vjta.2022.v03i01.03.
- [9] M. J. Balick and P. A. Cox, *Plants, people, and culture: the science of ethnobotany*, 2nd Edition, Garland Science, p. 228, 2020, doi: 10.4324/9781003049074.
- [10] A. Singh, V. Nath, S. Kumar, B. S. Singh, and B. Reddy, "The role of a traditional festival, chhath puja, in the conservation and sustainable use of tropical fruits," *Tropical Fruit Tree Diversity*, p. 217, 2016.
- [11] T. A. Shaikh, T. Rasool, and F. R. Lone, "Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming," *Computers and Electronics in Agriculture*, vol. 198, p. 107119, 2022, doi: 10.1016/j.compag.2022.107119.
- [12] P. Thongnim, V. Yuvanatemiya, and P. Srinil, "Smart agriculture: Transforming agriculture with technology," in *22nd Asia Simulation Conference*, Langkawi, Malaysia, Springer, 2023, pp. 362–376, doi: 10.1007/978-981-99-7240-1_29.
- [13] P. Thongnim, V. Yuvanatemiya, E. Charoenwanit, and P. Srinil, "Design and testing of spraying drones on durian farms," *2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, Jeju, Korea, Republic of, 2023, pp. 1–6, doi: 10.1109/ITC-CSCC58803.2023.10212524.
- [14] M. Hassan, K. Malhotra, and M. Firdaus, "Application of artificial intelligence in iot security for crop yield prediction," *ResearchBerg Review of Science and Technology*, vol. 2, no. 1, pp. 136–157, 2022.
- [15] E. Elbasi *et al.*, "Crop prediction model using machine learning algorithms," *Applied Sciences*, vol. 13, no. 16, p. 9288, 2023, doi: 10.3390/app13169288.
- [16] I. A. Abouelsaad, I. I. Teiba, E. H. El-Bilawy, and I. El-Sharkawy, "Artificial intelligence and reducing food waste during harvest and post-harvest processes," *IoT-Based Smart Waste Management for Environmental Sustainability*, CRC Press, 2022, pp. 63–82, doi: 10.1201/9781003184096-4.
- [17] V. K. Pandey *et al.*, "Machine learning algorithms and fundamentals as emerging safety tools in preservation of fruits and vegetables: a review," *Processes*, vol. 11, no. 6, p. 1720, 2023, doi: 10.3390/pr11061720.
- [18] J. Kos-Łabedowicz, "The issue of digital divide in rural areas of the european union," *Ekonomiczne Problemy Usług*, vol. 126, no. 1/2, pp. 195–204, 2017, doi: 10.18276/epu.2017.126/2-20.
- [19] V. Sharma, A. K. Tripathi, and H. Mittal, "Technological revolutions in smart farming: Current trends, challenges & future directions," *Computers and Electronics in Agriculture*, vol. 201, p. 107217, 2022, doi: 10.1016/j.compag.2022.107217.
- [20] P. Thongnim, E. Charoenwanit, and T. Phukseng, "Cluster quality in agriculture: Assessing gdp and harvest patterns in asia and europe with k-means and silhouette scores," *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, Kolkata, India, 2023, pp. 1–5, doi: 10.1109/IEMENTech60402.2023.10423469.
- [21] A. Et-taleby, M. Boussetta, and M. Benslimane, "Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image," *International Journal of Photoenergy*, vol. 2020, no. 1, p. 6617597, 2020, doi: 10.1155/2020/6617597.
- [22] M. Cui, "Introduction to the k-means clustering algorithm based on the elbow method," *Accounting, Auditing and Finance*, vol. 1, no. 1, pp. 5–8, 2020, doi: 10.23977/acaf.2020.010102.
- [23] J. Wala, H. Herman, R. Umar, and S. Suwanti, "Heart Disease Clustering Modeling Using a Combination of the K-Means Clustering Algorithm and the Elbow Method," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 903–914, 2024, doi: 10.15294/sji.v11i4.14096.
- [24] F. Ros, R. Riad, and S. Guillaume, "PdBi: A partitioning davies-bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, pp. 178–199, 2023, doi: 10.1016/j.neucom.2023.01.043.
- [25] C.-E. B. Ncir, A. Hamza, and W. Bouaguel, "Parallel and scalable dunn index for the validation of big data clusters," *Parallel Computing*, vol. 102, p. 102751, 2021, doi: 10.1016/j.parco.2021.102751.
- [26] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021, doi: 10.3390/e23060759.
- [27] L. Wilkinson, "Visualizing big data outliers through distributed aggregation," in *IEEE Transactions on Visualization and Computer Graphic*, vol. 24, no. 1, pp. 256–266, 2017, doi: 10.1109/TVCG.2017.2744685.

BIOGRAPHIES OF AUTHORS






Pattharaporn Thongnim    is an Assistant Professor at Burapha University in Thailand, holding a Ph.D. in statistics and data science from the University of Leicester, UK. Her career combines cutting-edge technology with data analysis, as she employs drones for innovative farming solutions and applies modern statistical and machine learning techniques to logistics, climate, and agricultural data. This multifaceted approach allows her to gain valuable insights into the complex relationship between environmental factors and their impact on agricultural economics and logistics. By bridging the gap between advanced data analysis and practical applications, her work contributes significantly to the fields of agriculture and technology. She can be contacted at email: pattharaporn@buu.ac.th.



Phaitoon Srinil    works as an Assistant Professor at Burapha University's Faculty of Science and Arts in Thailand. He got his B.Eng. and M.Eng. degrees in computer engineering from King Mongkut's Institute of Technology Ladkrabang (KMITL) in Bangkok. Now, he teaches in the Department of Information Technology and Data Science (ITDS) at Burapha University in Chanthaburi, Thailand. He is interested in many areas of research such as the internet of things (IoT) and its use in farming, recognizing patterns, deep learning, smart technology, and the growing field of artificial intelligence (AI). He can be contacted at email: phaitoon@buu.ac.th.



Thanaphon Phukseng    is an Assistant Professor at the Faculty of Science and Arts, Burapha University Chanthaburi Campus, Thailand. He graduated with a B.Sc. in computer information system from Burapha University, a M.Sc. in information technology from King Mongkut's Institute of Technology Ladkrabang (KMITL) and a Ph.D. in information technology from King Mongkut's University of Technology North Bangkok, Thailand. Currently, he is a dedicated lecturer in the Department of Information Technology and Data Science (ITDS) at the Faculty of Science and Arts, Burapha University Chanthaburi Campus, Chanthaburi Province, Thailand. His research spans a broad array of interests, including recommendation systems, electronic commerce, graph theory, and data visualization. He can be contacted at email: thanaph@buu.ac.th.